

The PowerStack Initiative

A Community-driven Effort

EEHPC-WG Webinar Series
September 12, 2018

PowerStack Core Committee (alphabetical order)

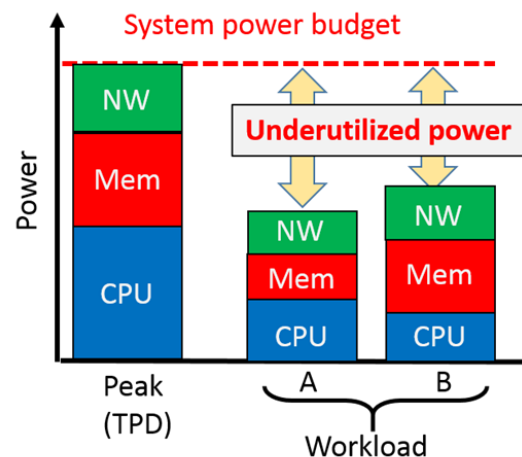
- Cantalupo, Christopher (Intel, USA)
- Eastep, Jonathan (Intel, USA)
- **Jana, Siddhartha (EEHPC WG, USA) <-- Speaker**
- Kondo, Masaaki (Univ of Tokyo, Japan)
- Maiterth, Matthias (LMU Munich, Germany)
- Marathe, Aniruddha (LLNL, USA)
- Patki, Tapasya (LLNL, USA)
- Rountree, Barry (LLNL, USA)
- Sakamoto, Ryuichi (Univ of Tokyo, Japan)
- Schulz, Martin (TU-Munich, Germany)
- Trinitis, Carsten (TU-Munich, Germany)

Outline

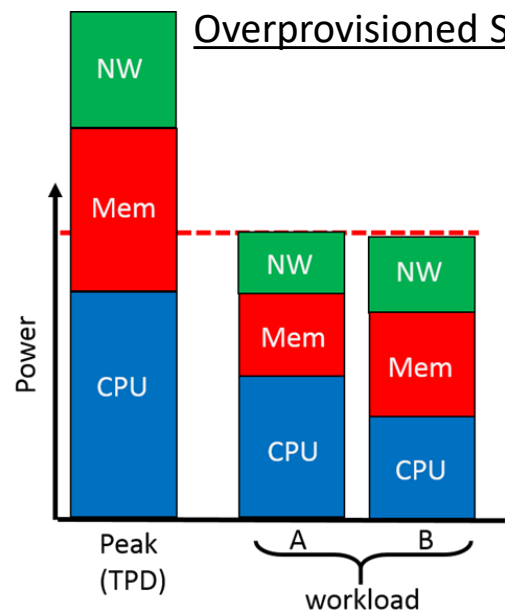
- Motivation
- Charter of the PowerStack initiative
- Stakeholders and research collaborators
- PowerStack overview
 - PowerStack block diagram
 - Existing Components
 - Working groups
- Next steps / Call to Action

Resource Utilization Variability at System-Level

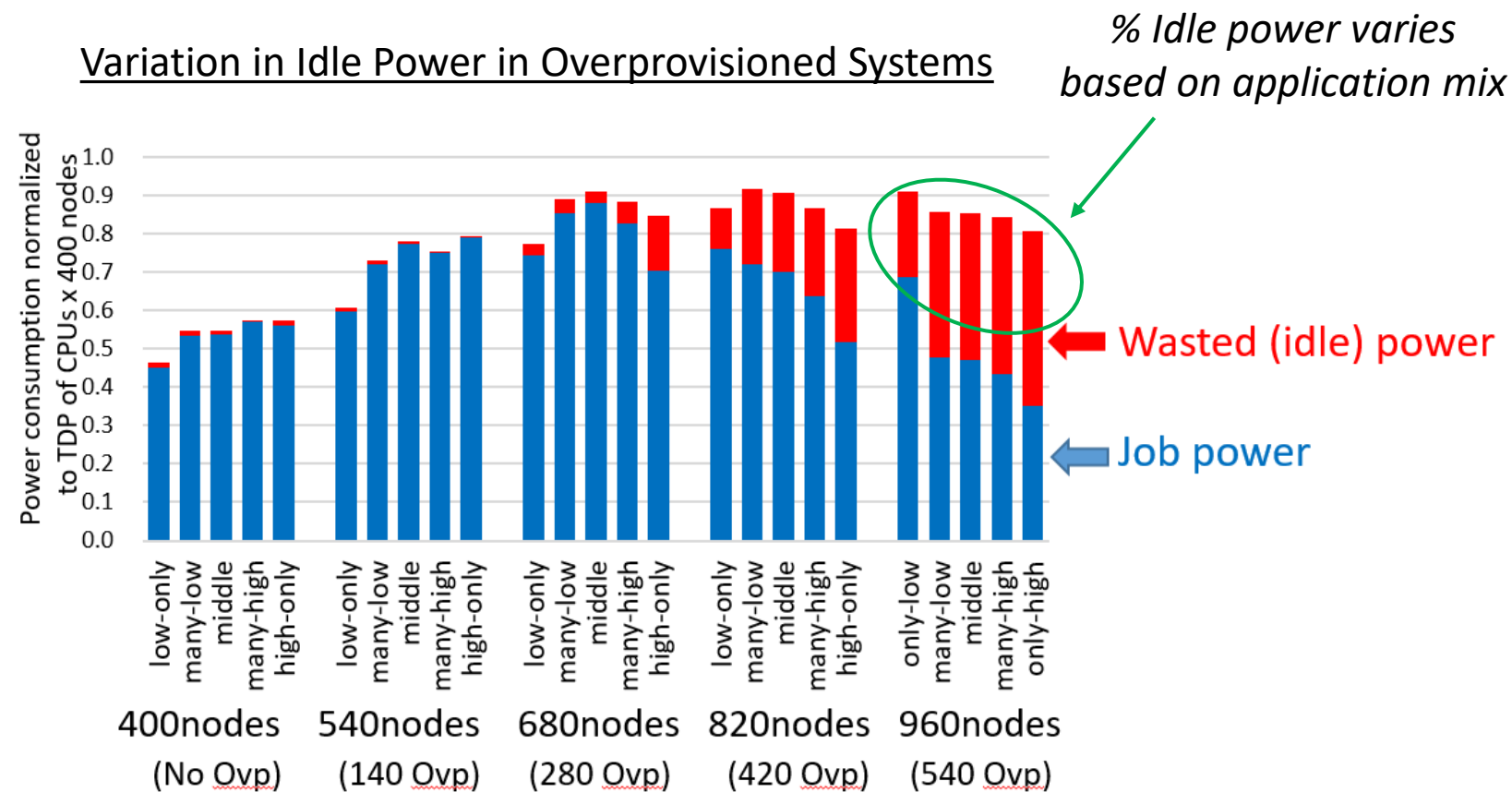
Conventional Systems



Overprovisioned Systems



Variation in Idle Power in Overprovisioned Systems



Source: IPDPS 2018, Proceedings, Sakamoto et al., "Analyzing Resource Trade-offs in Hardware Overprovisioned Supercomputers"

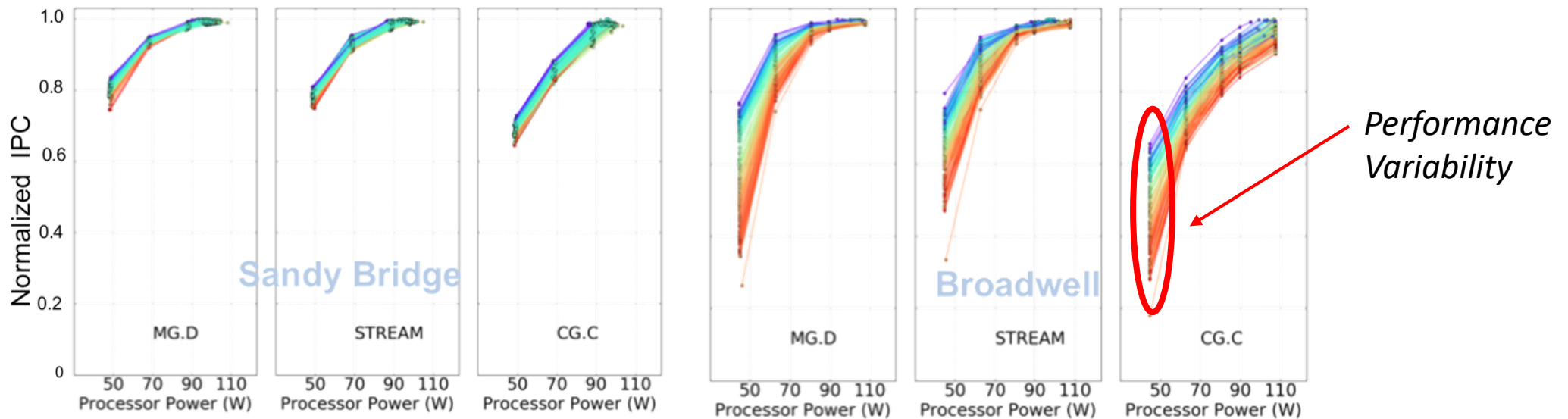
Performance Variability at Job- and Platform-Level

❖ Causes

1. **Application design – bulk synchronization, collective operations**
2. **Non-deterministic topology due to node availability**
3. **Manufacturing variability**
 - Performance differences no longer compensates for power consumption
 - Continues to increase, and will worsen with heterogeneity
 - 4x difference between two generations of Intel processors

❖ Needs advanced runtime options for mitigation

- Need to know power/performance profile of each socket
- Average power caps will create load imbalance



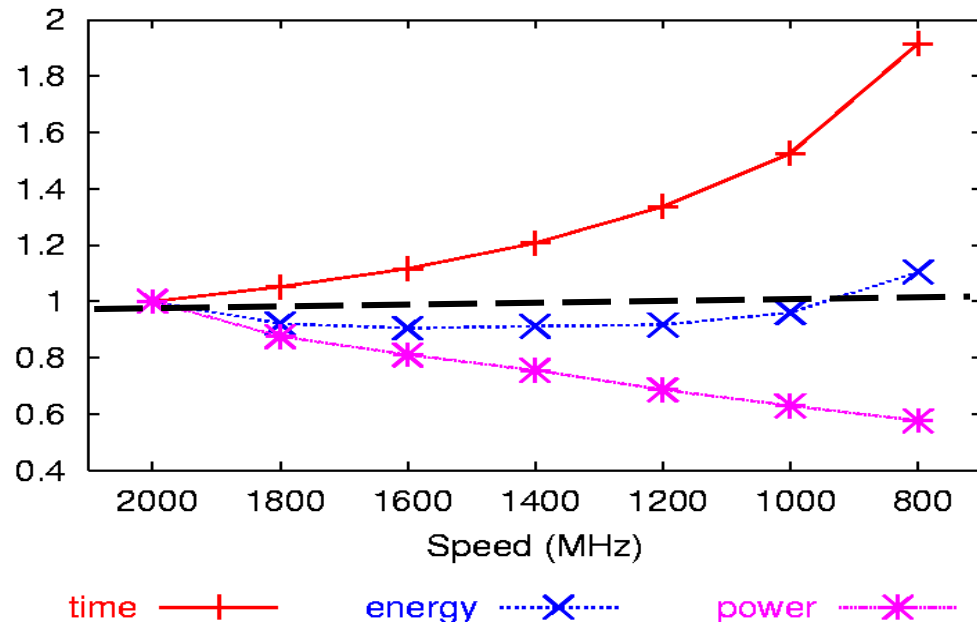
Source: ISC 2018, Tutorials, Schulz et al., "Boosting Power Efficiency of HPC Applications with GEOPM"

Race-to-halt is not a solution!

Impact of CPU Frequency Scaling of a NAS kernel

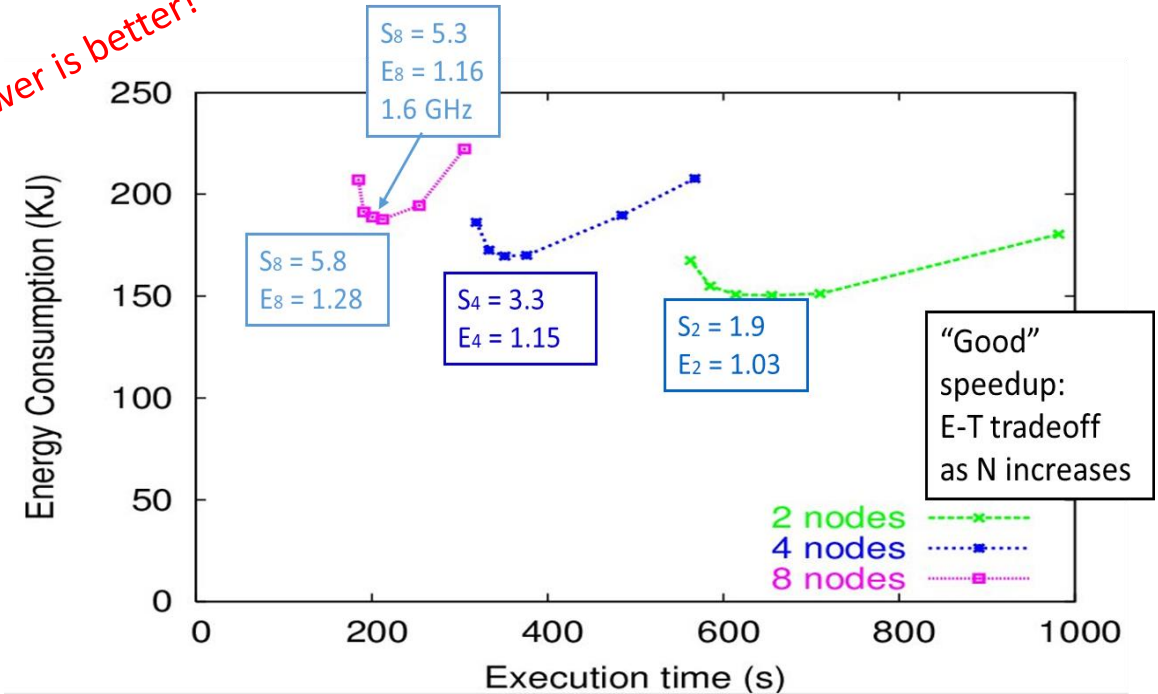
Single Node

Lower is better!



Multiple Nodes

Lower is better!



Takeaway:

- Extent of Speedup gains and Energy savings due to DVFS drops with N
- Behavior of application varies with N

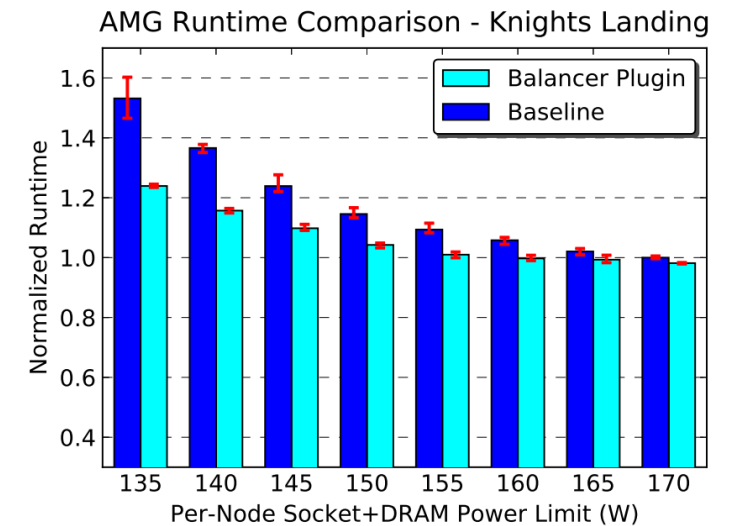
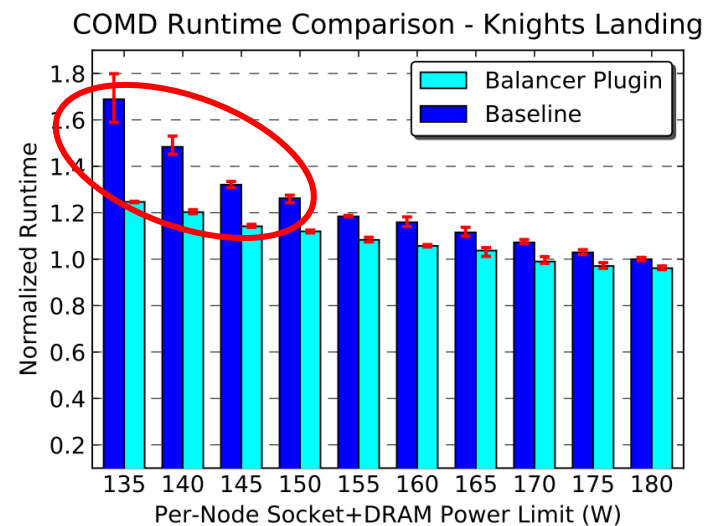
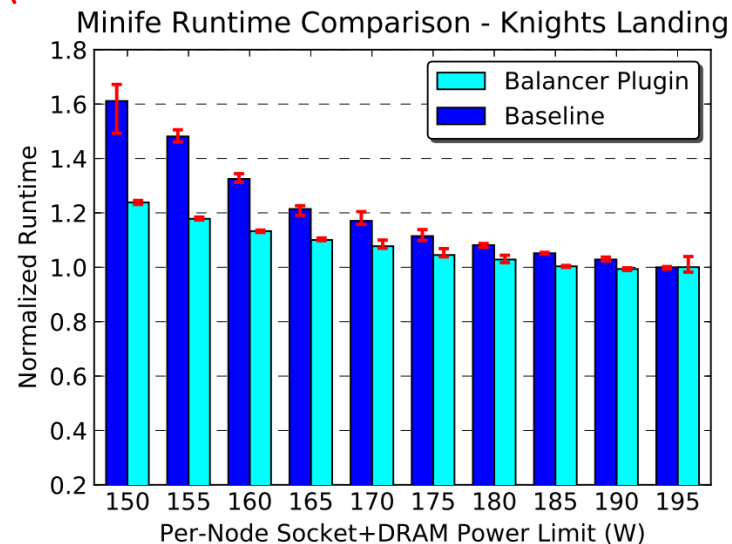
Source: ISC 2018, Tutorials, Schulz et al., "Boosting Power Efficiency of HPC Applications with GEOPM"

Introducing Job-awareness within HPC Systems

- ❖ State-of-the-art application-level runtime systems help cover the efficiency gap!
- ❖ Leverage job-awareness while driving system-wide efficiency
 - E.g. Use of GEOPM, free open source hierarchical distributed runtime
- ❖ Provides application-aware dynamic optimization of HW power knob settings
- ❖ Up to 30% reduction in application time-to-solution in power-capped systems



Lower is better!



Source: ISC 2017, Proceedings, Eastep et al., "GEOPM: A vehicle for HPC community collaboration on co-designed energy management solutions"

The PowerStack Initiative

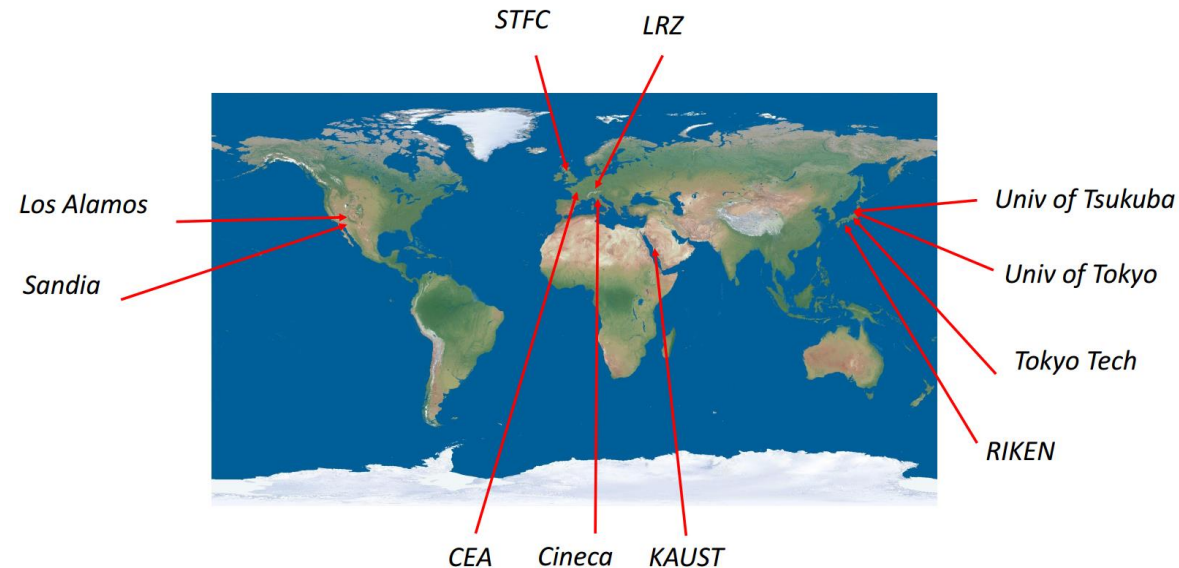
- ❖ Collaboration towards a well-defined, community-wide stack that accounts for power-awareness across various layers of the HPC *software* ecosystem
- ❖ Charter:
 1. Identify **different actors** that play a role in energy- and power-aware job scheduling and resource mgmt
 2. Reach a community-wide consensus on the **roles and responsibilities** of the different actors, their **interoperability**, and communication **protocols**
 3. Work towards prototypes and full-scale production-grade solutions that are **adaptive and feedback-driven**

PowerStack Stakeholders

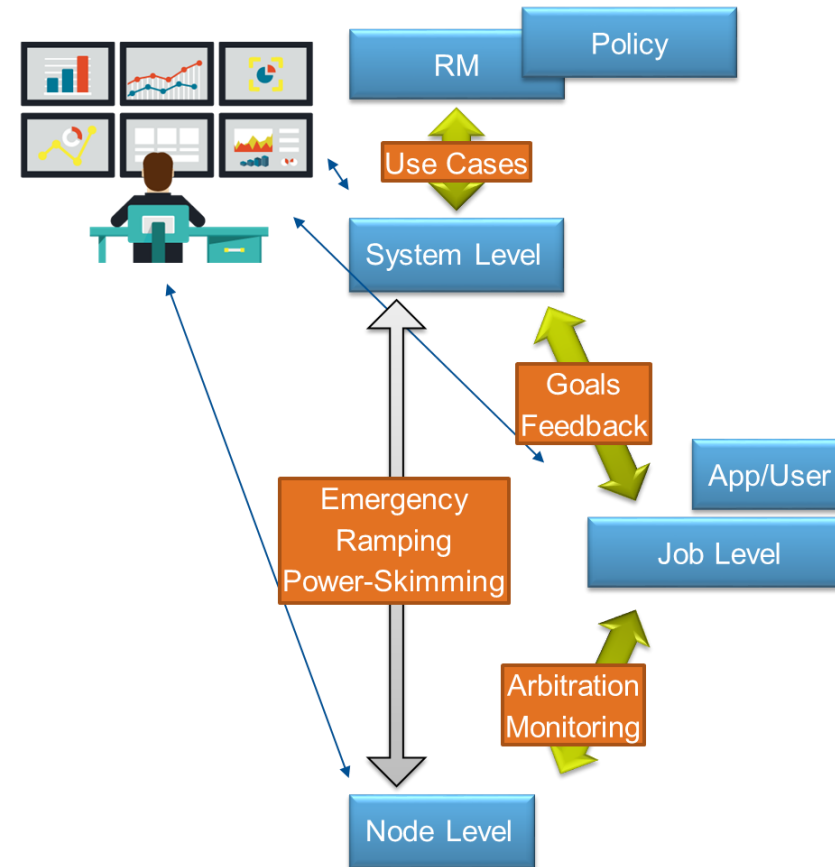
Participants of the PowerStack seminar (June 2018):

- ❖ LLNL, LANL, Sandia, Argonne, Riken, STFC
- ❖ ATOS/Bull, Cray, Fujitsu, IBM, Intel, AMD, ARM, HPE, Altair
- ❖ TU-Munich, TU-Dresden, UniBo, SDU, Univ of Tokyo, LRZ, Grenoble, EEHPC-WG

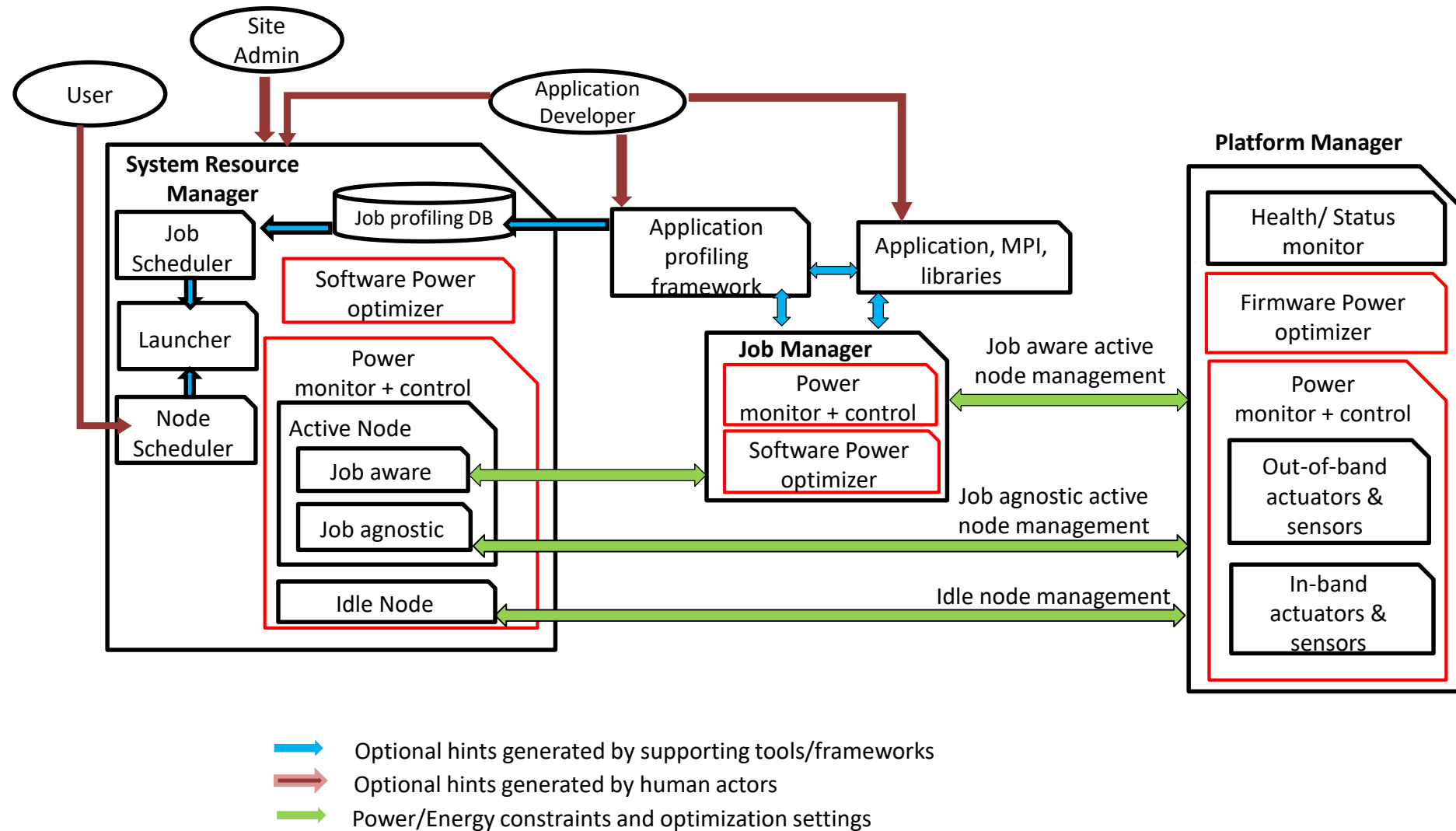
EEHPC-WG's insight into sites investing in Energy- and Power-aware Job Scheduling and Resource Management (*EPA-JSRM*)



PowerStack - Layers



PowerStack - 3 key actors



Examples of PowerStack components

| PowerStack Software actors | Examples of current state-of-the-art components |
|-----------------------------------|--|
| Workload Management | Slurm, ALPS, PBSPro, Cobalt |
| Application / Job manager | GEOPM, Conductor |
| Platform / Node manager | PAPI, PowerAPI, Variorum, NVML, Redfish, HDEEM, Application runtime params, Fabric manager |

Working Groups catering to specific topics

Working Group 1:

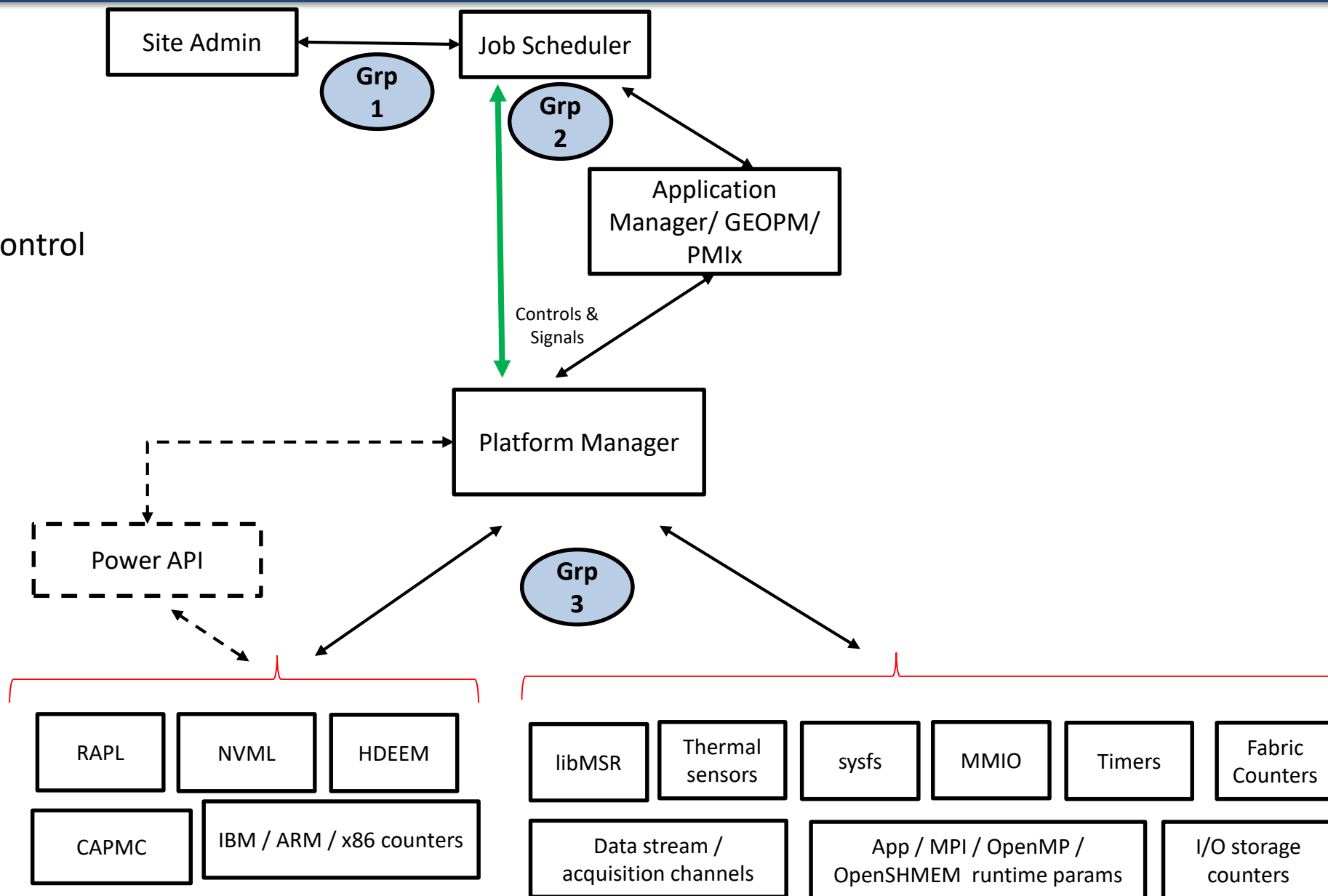
- PowerStack Site policy

Working Group 2:

- PowerStack Adaptive/Runtime control

Working Group 3:

- PowerStack platform interaction



Call to Action

Next Steps:

1. First General Membership Meeting between Sept 17-28, 2018
2. Topic-specific working-groups specific Periodic Meetings
 - Once every 1.5 months (~ 6 weeks)
 - Subscribe to Mailing lists

Mailing list names:

- PowerStack Announcements powerstack-announce@googlegroups.com
- PowerStack Development powerstack-dev@googlegroups.com
- PowerStack Adaptive Runtime and Control powerstack-runtime@googlegroups.com
- PowerStack Platform Interaction powerstack-platform@googlegroups.com
- PowerStack Site Policy and Verification powerstack-sitepolicy@googlegroups.com

Thanks!

Acknowledgements:

- PowerStack Core Committee
- PowerStack Seminar Attendees

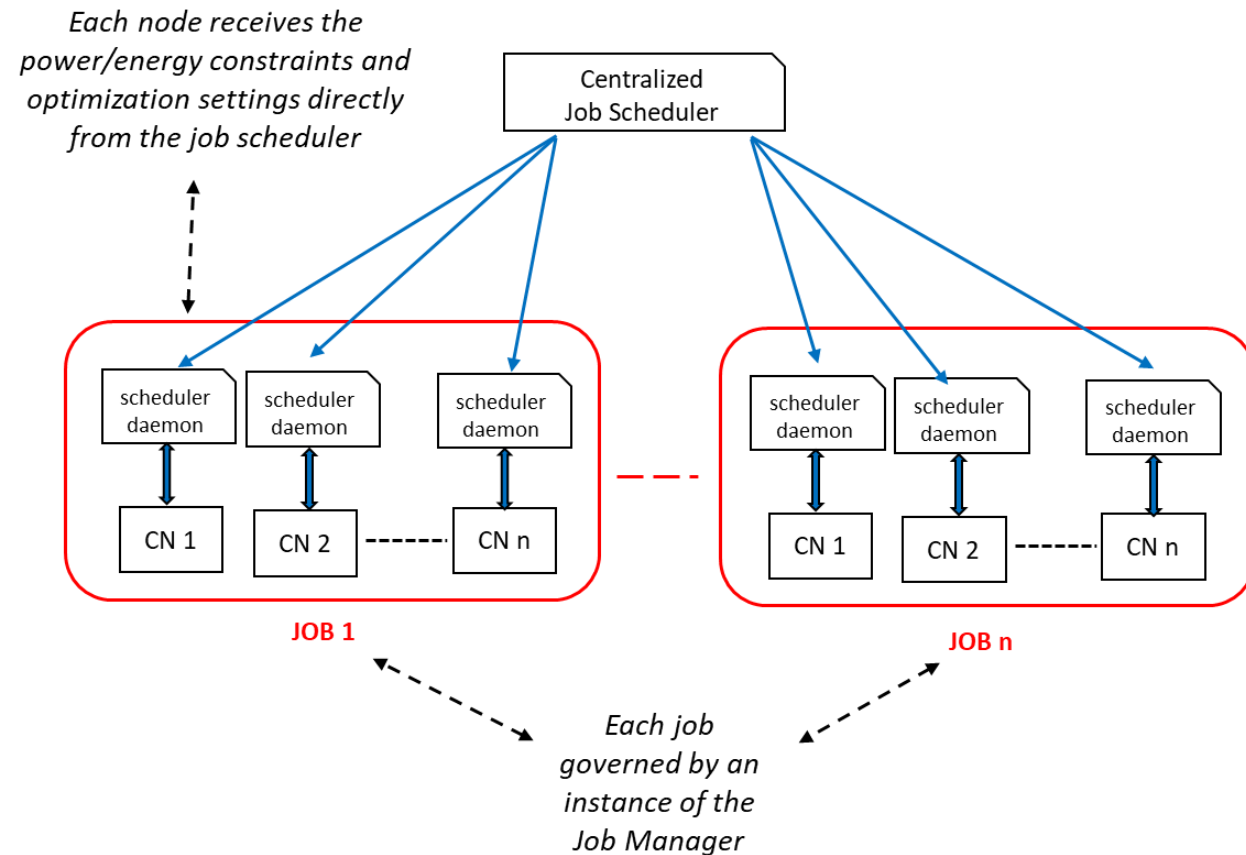
Speaker Contact:

- Sid Jana siddhartha.jana@intel.com

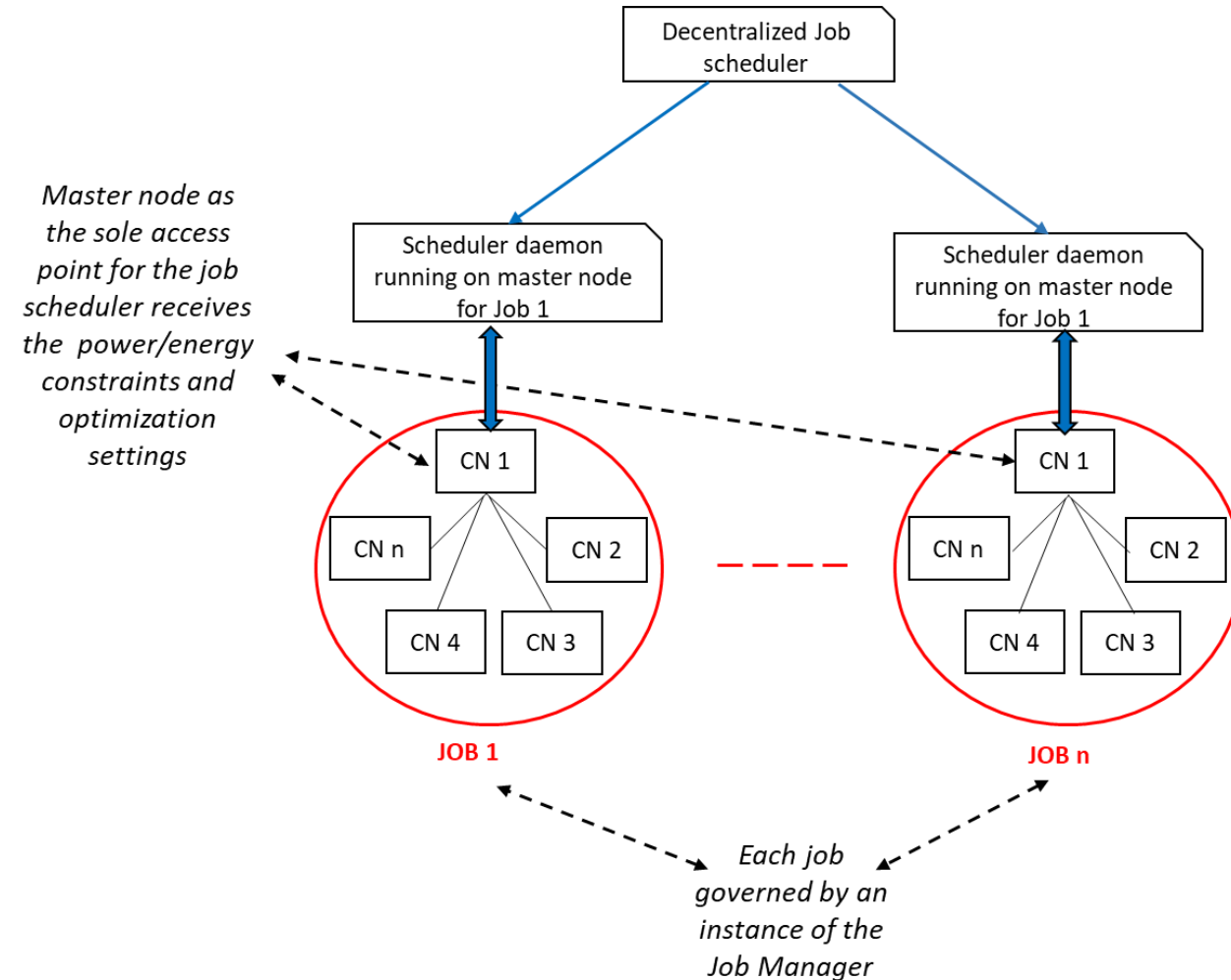
Backup Slides

Interoperability between a Job Scheduler and a Job Manager

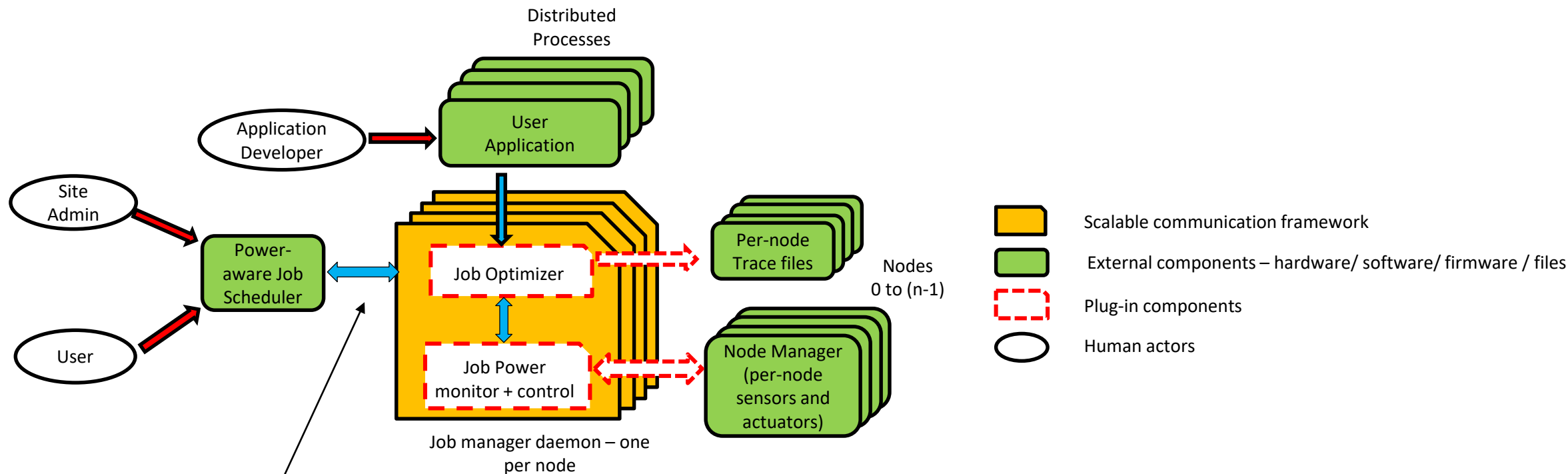
Current Approach (centralized)



Proposed Approach (decentralized) (pending community consensus)



Proposed Design of a Job Manager



*Community-wide discussions
on bidirectional
communication channel
commencing soon!*

Example of an open-source Job Manager: GEOPM

- Globally Extensible Open Power Manager
- Included within OpenHPC
- Open-source (BSD license), platform/vendor agnostic
- <https://geopm.github.io/>
- W.I.P. deployments: Theta (Argonne), Cori (NERSC), Trinity (LANL/Sandia), Quartz (LLNL), SuperMUC-NG (LRZ)